­­­­­CSCI-572 Homework 4: Indexing the Web Using Solr

Name: Maria Divya Fernandes

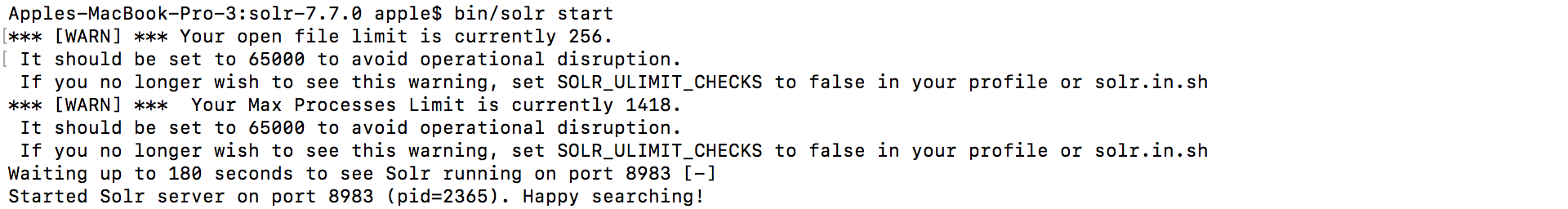
USC ID :8350120470

Website Indexed – Reuters

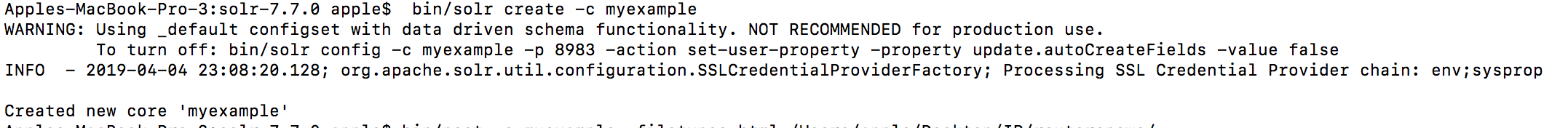
Step 1: Installing Solr

* I downloaded solr from <http://archive.apache.org/dist/lucene/solr/7.7.0/>
* Open the Terminal > cd into solr-7.x.x folder and enter the command: bin/solr start





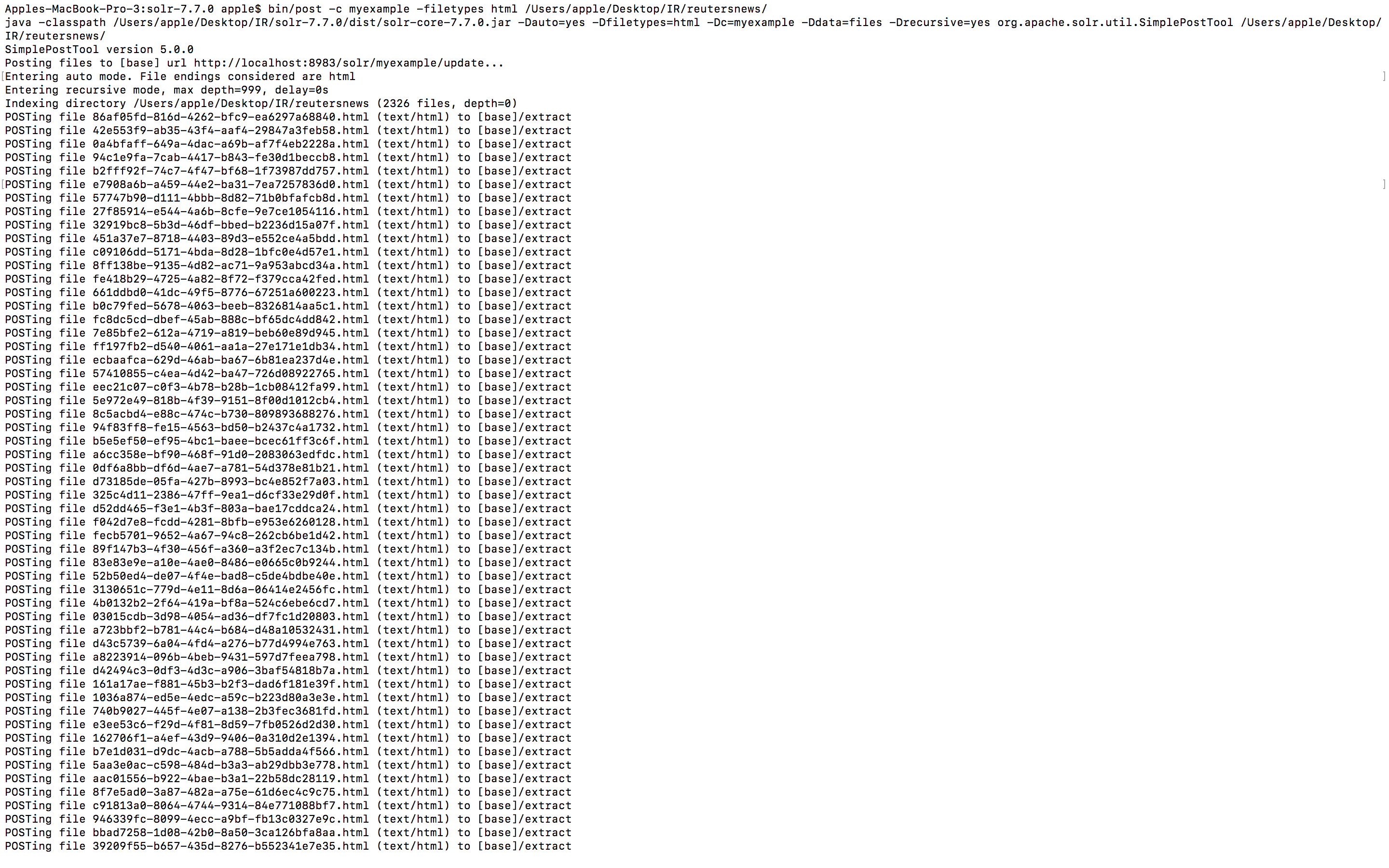
* Create a new core using the command: bin/solr create –c myexample

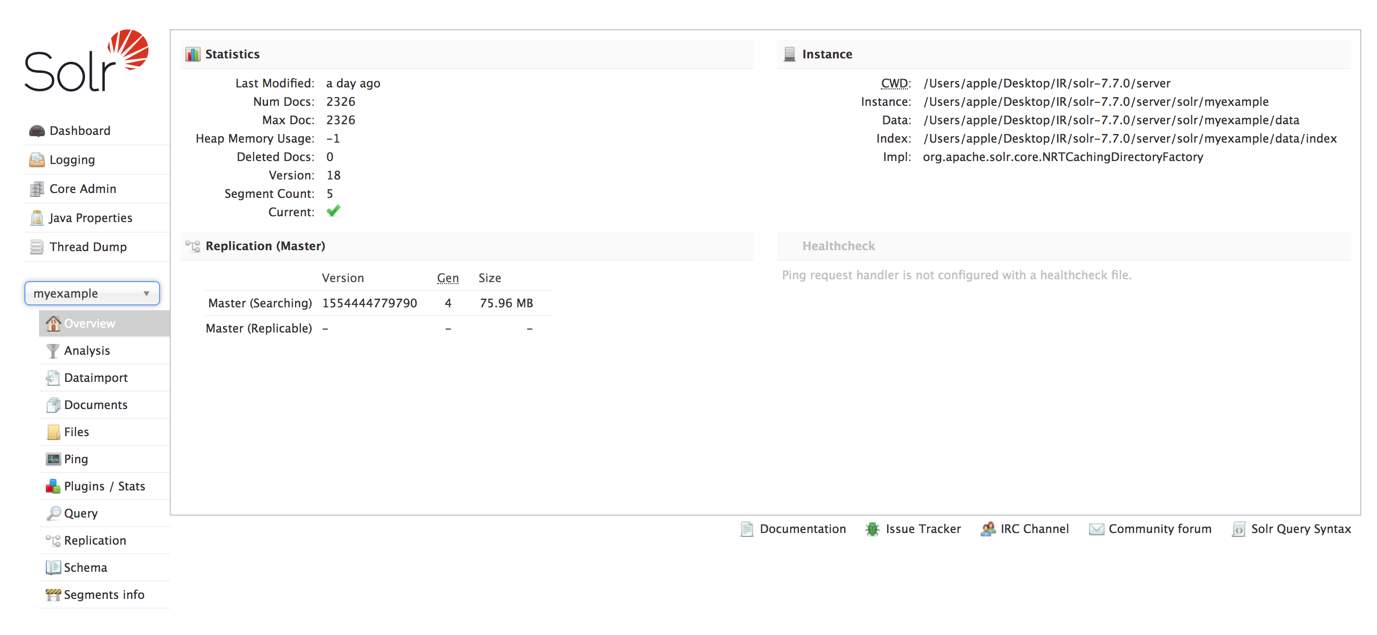


* Solr was seen on the web browser at port 8983 – [http://localhost:8983/solr/#/](http://localhost:8983/solr/#/myexample/query)

Step 2: Indexing The Web Page

* Download the data (reutersnews and URLtoHTML\_reuters\_news.csv) from the google drive <https://drive.google.com/drive/u/0/folders/1EFkdy9cLjq0hNuC_bRQ0UO75djtvwrj7?usp=sharing> for Reuters .
* Indexing is performed by using the command: bin/post -c myexample -filetypes html /Users/apple/Desktop/IR/reutersnews/

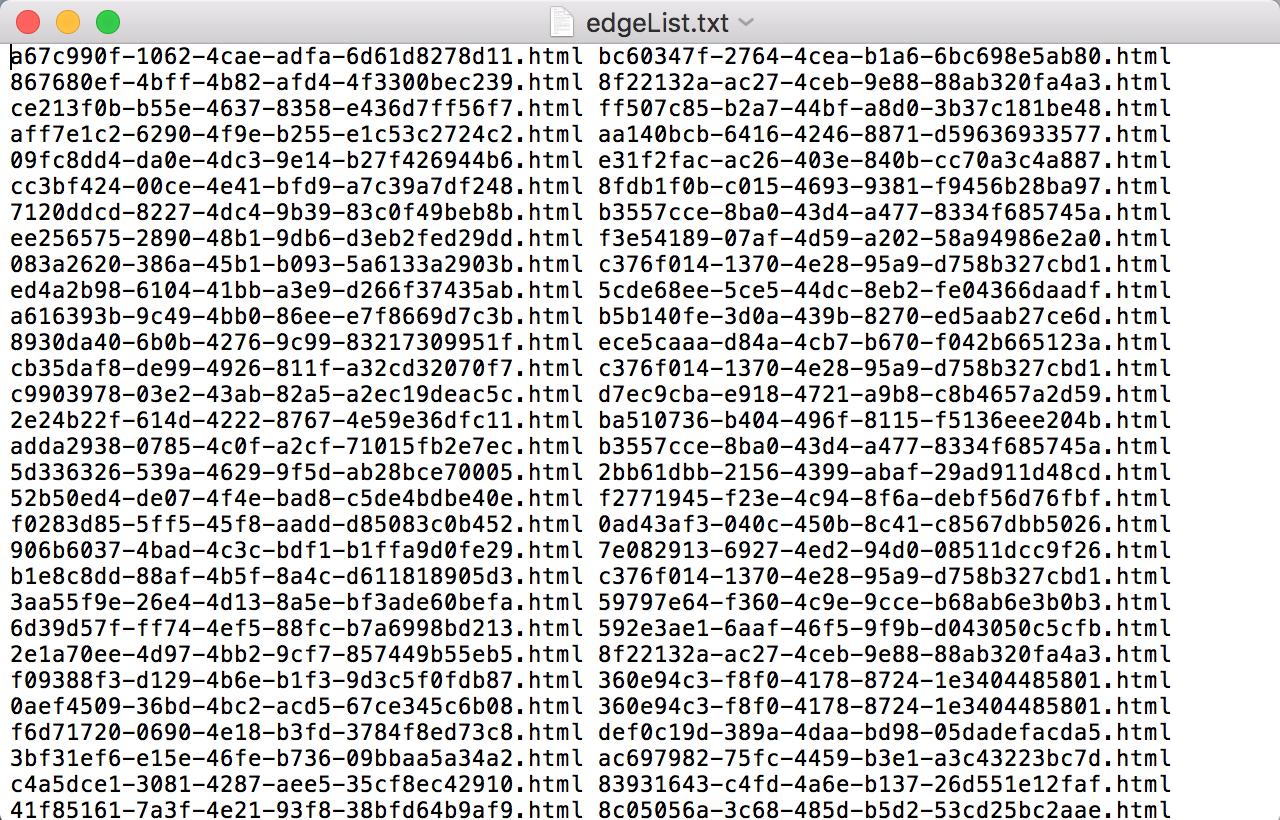




Step 3 : Implementing PageRank Algorithm:

* I downloaded jsoup-1.11.3.jar from the <https://jsoup.org/download> and wrote a Java file ListLink.java to find the edgelist for the Reuters.

The following is the edge list snippet



* The I used Networkx to help you compute the PageRank given a graph to read the graph and output a pagerank file name external\_pageRankFile.txt as given in the folder.
* PageRank Factors : alpha=0.85, personalization=None, max\_iter=30, tol=1e-06, nstart=None, weight='weight',dangling=None
* Then I copied the file into the data folder of my core (myexample).

I executed the php code using XAMPP by placing maria.php and solr-php-client-master inside htdocs and entering localhost/maria.php in the browser.

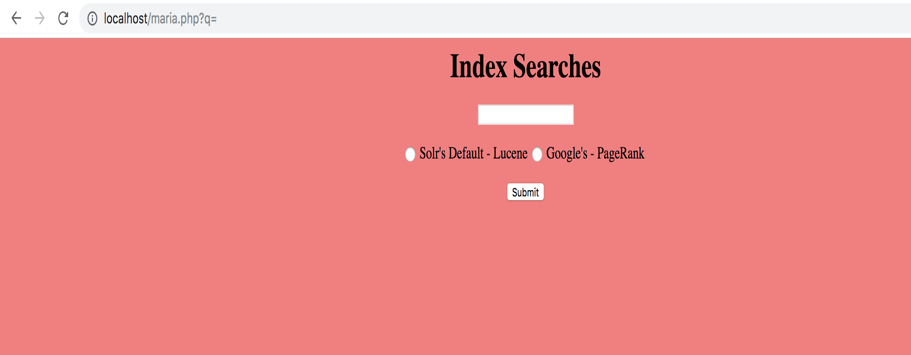
When running the php code kindly change the path of the solr-php-client-master file

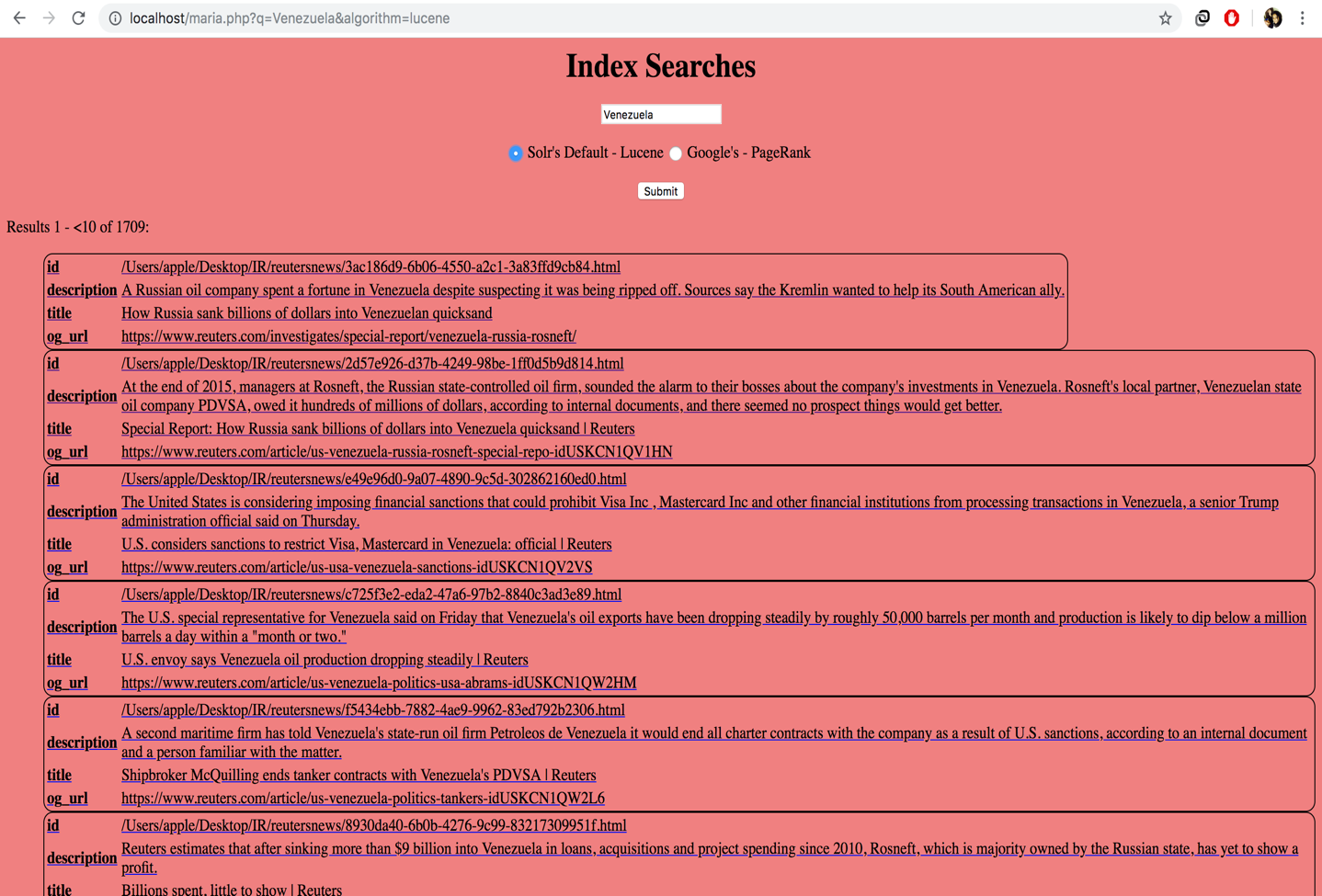
Line 10 : require\_once('/Applications/XAMPP/xamppfiles/htdocs/solr-php-client-master/Apache/Solr/Service.php');

**The Following Snippet Shows a working model of my code :**

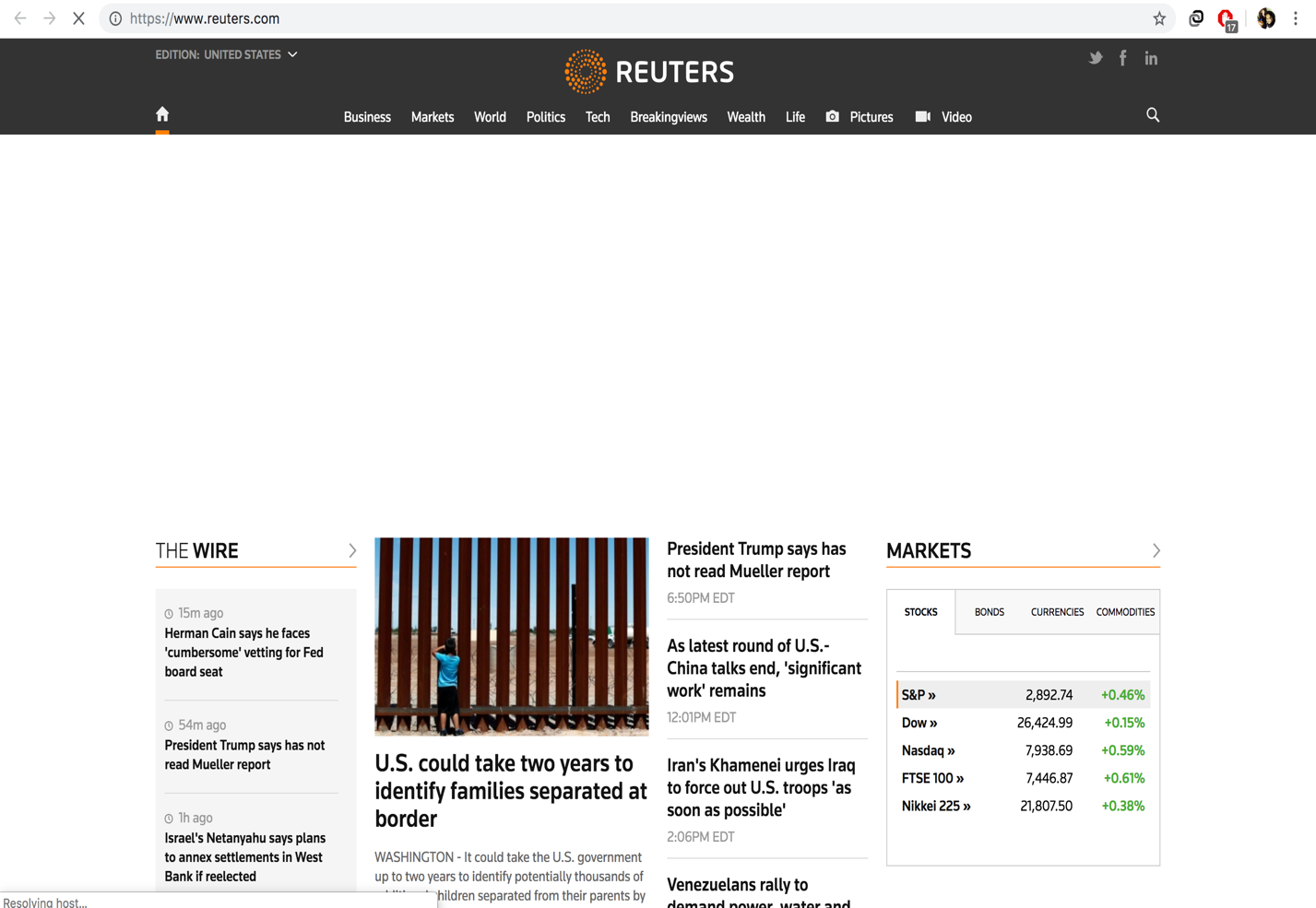
1. Steps in Query

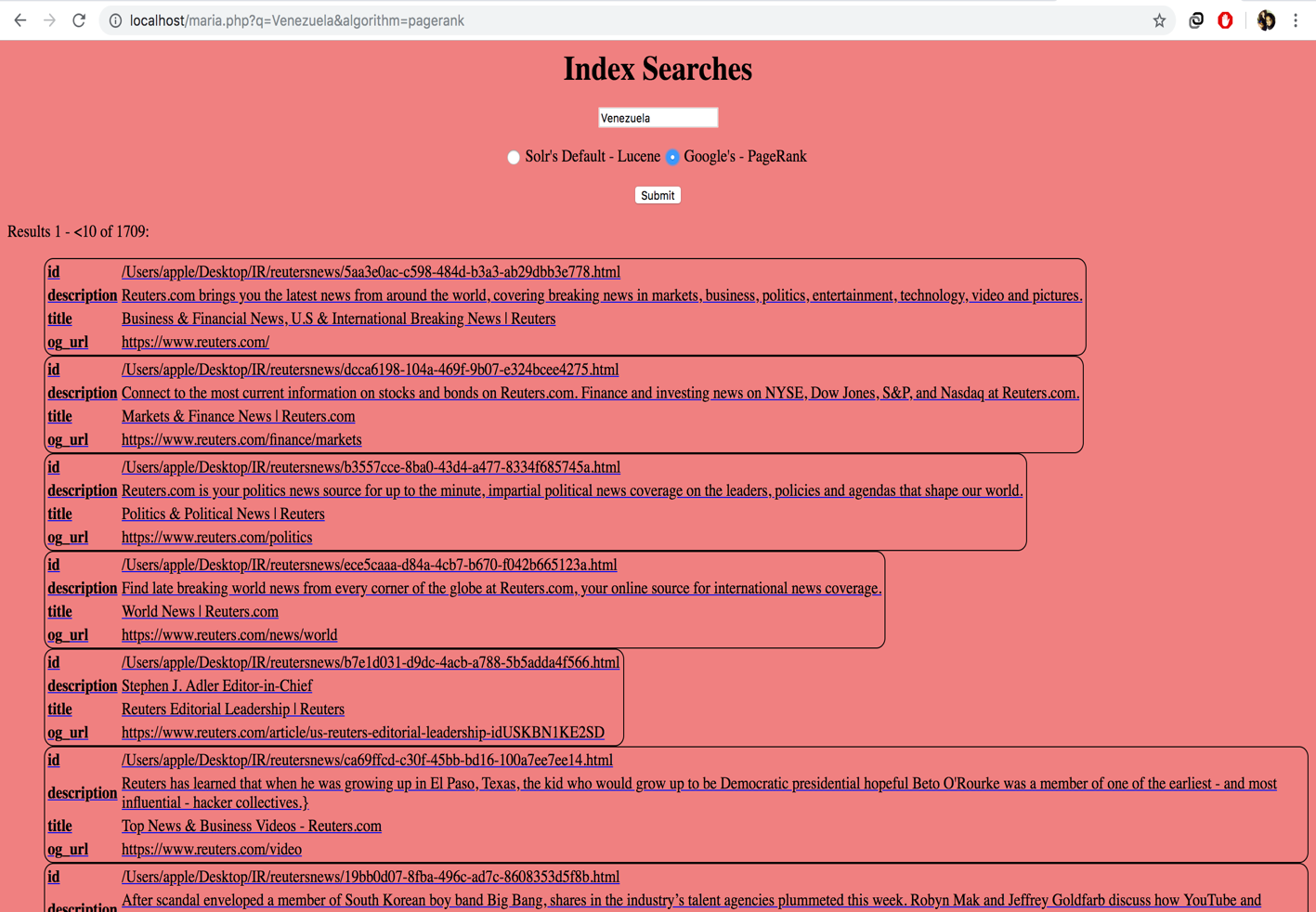
maria.php gives the user interface on the localhost for querying solr. We can choose the algorithm to be executed – Lucene or PageRank. The results are then displayed on the same page. These results are clickable links that can be routed to the destination link. The following snippets shows the process of Querying solr.

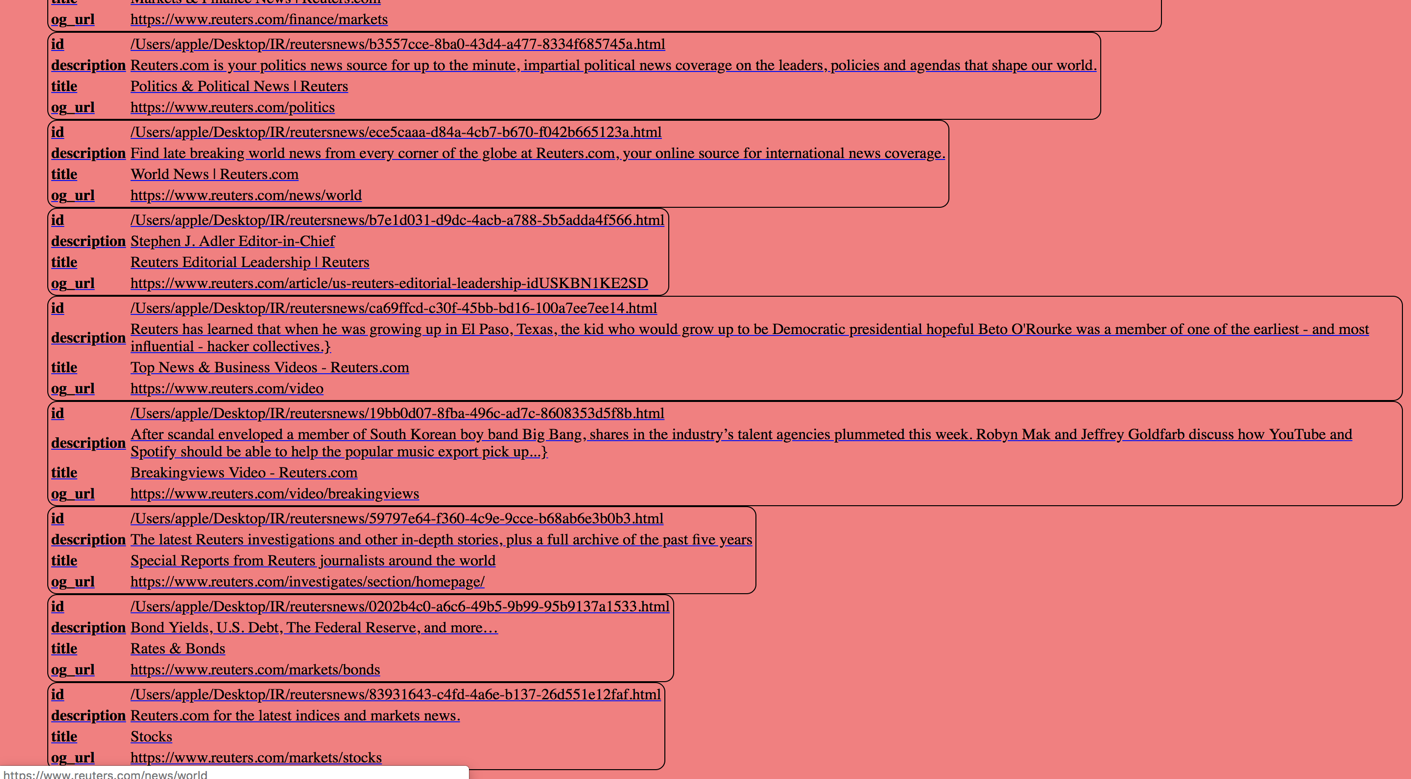


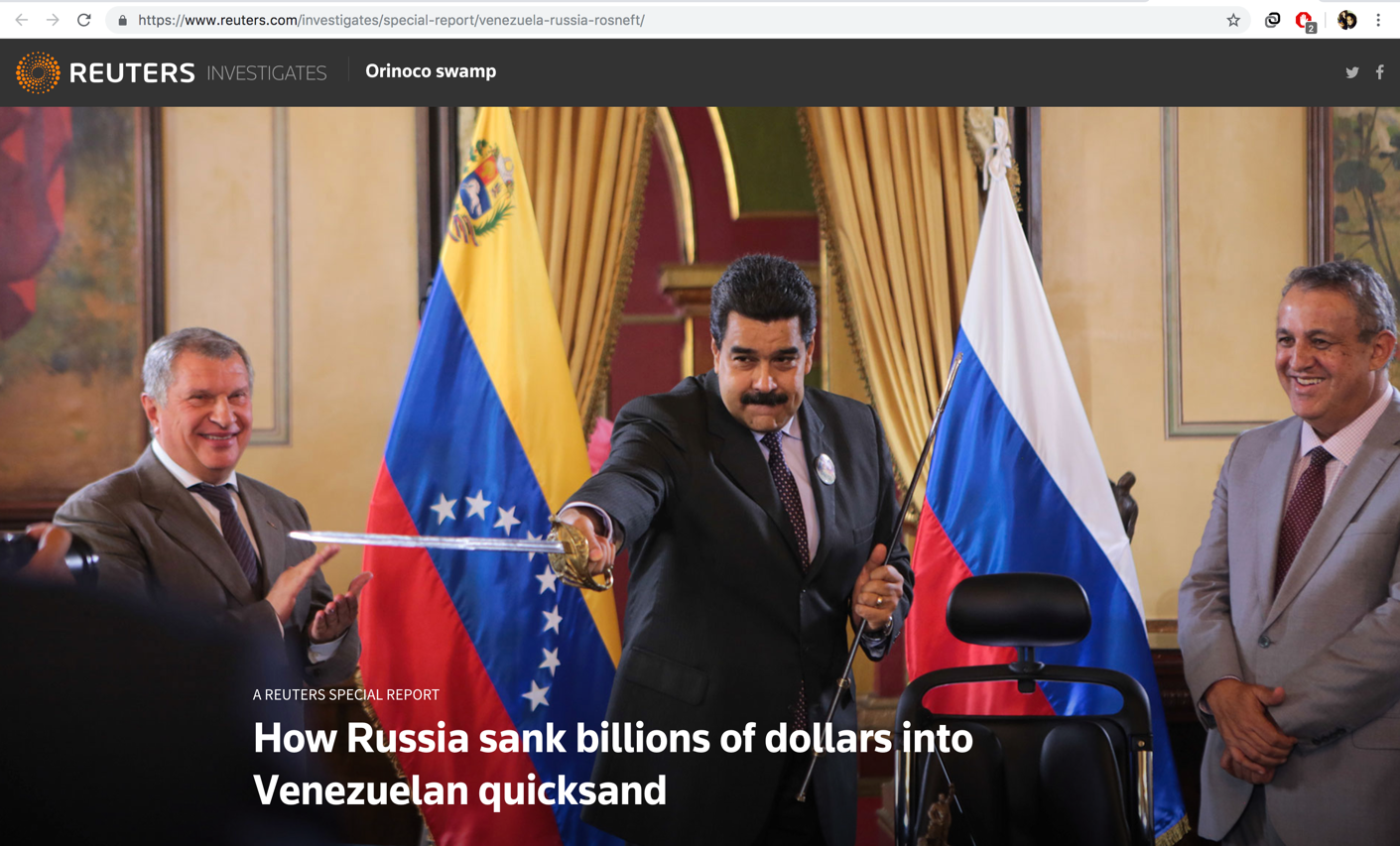












**Explanation regarding why some pages have higher PageRank values than others.**

PageRank calculates weight for a page based on the number of links pointing to that page. We create a graph that determines the list of edges pointing to the page. This graph is then used to find the importance of a given page in the web. Solr algorithm uses only the vector and boolean model that just scans the tf-idf value and outputs the results. However, PageRank determines the weight of the page and then outputs it in a required order. Thus, the results follow a particular ranking and are different .Thus the major reason that pagerank has higher page rank value is the fact that it analyses the weight of page on the number of links that point it.

Query 1: Venezuela

|  |  |
| --- | --- |
| Lucene | PageRank |
| <https://www.reuters.com/investigates/special-report/venezuela-russia-rosneft/> | https://www.reuters.com/ |
| https://www.reuters.com/article/us-venezuela-russia-rosneft-special-repo-idUSKCN1QV1HN | https://www.reuters.com/finance/markets |
| https://www.reuters.com/article/us-usa-venezuela-sanctions-idUSKCN1QV2VS | https://www.reuters.com/politics |
| https://www.reuters.com/article/us-venezuela-politics-usa-abrams-idUSKCN1QW2HM | https://www.reuters.com/news/world |
| https://www.reuters.com/article/us-venezuela-politics-tankers-idUSKCN1QW2L6 | https://www.reuters.com/article/us-reuters-editorial-leadership-idUSKBN1KE2SD |
| https://www.reuters.com/article/us-venezuela-russia-rosneft-money-idUSKCN1QV1HS | https://www.reuters.com/video |
| https://www.reuters.com/journalists/nelson-bocanegra | https://www.reuters.com/video/breakingviews |
| https://www.reuters.com/journalists/lesley-wroughton | https://www.reuters.com/investigates/section/homepage/ |
| https://www.reuters.com/journalists/roberta-rampton | https://www.reuters.com/markets/bonds |
| https://www.reuters.com/article/us-venezuela-politics-usa-banks-idUSKCN1QN2QE | https://www.reuters.com/markets/stocks |

Query 2: Senate

|  |  |
| --- | --- |
| Lucene | PageRank |
| <https://www.reuters.com/article/us-usa-trump-congress-emergency-idUSKCN1QV1A8> | https://www.reuters.com/politics |
| https://www.reuters.com/video/2019/03/04/senate-has-votes-to-block-border-emergen?videoId=521958334&videoChannel=-10465 | https://www.reuters.com/video |
| https://www.reuters.com/video/2019/03/05/mcconnell-senate-has-votes-to-block-bord?videoId=521939215&videoChannel=-10465 | https://www.reuters.com/video/breakingviews |
| https://www.reuters.com/article/us-usa-politics-beto-orourke-special-rep-idUSKCN1QW26N | https://www.reuters.com/subjects/sustainable-business |
| https://www.reuters.com/article/us-usa-politics-beto-orourke-special-rep-idUSKCN1QW26N | https://www.reuters.com/news/media |
| https://www.reuters.com/article/us-usa-politics-beto-short-exclusive-idUSKCN1QW28H | https://www.reuters.com/places/mexico |
| https://www.reuters.com/article/us-usa-politics-beto-short-exclusive-idUSKCN1QW28H | https://www.reuters.com/video/2019/03/14/orourke-hits-the-ground-running-in-iowa?videoId=525923095&videoChannel=1003 |
| <https://www.reuters.com/politics> | https://www.reuters.com/article/us-usa-trump-congress-emergency-idUSKCN1QW2AK |
| https://www.reuters.com/article/us-congo-politics-idUSKCN1QW28R | https://www.reuters.com/article/us-usa-politics-beto-short-exclusive-idUSKCN1QW28H |
| https://www.reuters.com/article/us-usa-election-orourke-idUSKCN1QV00E | https://www.reuters.com/article/us-usa-election-orourke-idUSKCN1QV13C |

Query 3: Democrats

|  |  |
| --- | --- |
| Lucene | PageRank |
| <https://www>.reuters.com/article/us-usa-election-medicare-idUSKBN1QU189 | <https://www>.reuters.com/ |
| <https://www>.reuters.com/article/us-usa-election-medicare-idUSKBN1QU189 | <https://www>.reuters.com/politics |
| <https://www>.reuters.com/article/us-usa-congress-whitaker-idUSKCN1PX1LD | <https://www>.reuters.com/subjects/supreme-court |
| <https://www>.reuters.com/video/2019/02/28/republicans-bash-cohen-democrats-at-cpac?videoId=520293110&videoChannel=13976 | <https://www>.reuters.com/breakingviews |
| <https://www>.reuters.com/article/us-usa-election-biden-idUSKCN1QH1EX | <https://www>.reuters.com/news/archive/politicsNews |
| <https://www>.reuters.com/article/us-usa-trump-budget-idUSKBN1QS12D | <https://www>.reuters.com/article/us-usa-trump-congress-emergency-idUSKCN1QW2AK |
| <https://www>.reuters.com/video/2019/03/05/the-witch-hunt-continues-trump-on-democr?videoId=522293719&videoChannel=13976 | <https://www>.reuters.com/article/us-usa-election-sanders-idUSKCN1QW2ZJ |
| <https://www>.reuters.com/article/us-usa-trump-russia-congress-idUSKCN1QV23E | <https://www>.reuters.com/article/us-usa-politics-beto-short-exclusive-idUSKCN1QW28H |
| <https://www>.reuters.com/politics | <https://www>.reuters.com/article/us-usa-tax-survey-idUSKCN1QW1BY |
| <https://www>.reuters.com/article/us-usa-trump-congress-emergency-idUSKCN1QW2AK | <https://www>.reuters.com/article/us-usa-election-booker-idUSKCN1QW2P4 |

Query 4: Republicans

|  |  |
| --- | --- |
| Lucene | PageRank |
| https://www.reuters.com/article/us-usa-trump-congress-emergency-idUSKCN1QV1A8 | https://www.reuters.com/politics |
| https://www.reuters.com/video/2019/02/28/republicans-bash-cohen-democrats-at-cpac?videoId=520293110&videoChannel=13976 | https://www.reuters.com/news/us |
| https://www.reuters.com/article/us-usa-trump-congress-emergency-idUSKCN1QW2AK | https://www.reuters.com/news/archive/politicsNews |
| https://www.reuters.com/article/us-usa-trump-congress-emergency-idUSKCN1QW2AK | https://www.reuters.com/article/us-usa-trump-congress-emergency-idUSKCN1QW2AK |
| https://www.reuters.com/article/us-usa-trump-congress-emergency-idUSKCN1QW2AK | https://www.reuters.com/article/us-usa-tax-survey-idUSKCN1QW1BY |
| https://www.reuters.com/article/us-usa-trump-congress-emergency-idUSKCN1QW2AK | https://www.reuters.com/article/us-usa-election-orourke-idUSKCN1QV13C |
| https://www.reuters.com/article/us-california-fires-trump-idUSKCN1P31ND | https://www.reuters.com/article/us-usa-immigration-wall-insight-idUSKCN1QW1GR |
| https://www.reuters.com/article/us-usa-tax-survey-idUSKCN1QW1BYhttps://www.reuters.com/article/us-usa-tax-survey-idUSKCN1QW1BY | https://www.reuters.com/article/us-usa-trump-russia-congress-idUSKCN1QV23E |
| https://www.reuters.com/article/us-usa-tax-survey-idUSKCN1QW1BY | https://www.reuters.com/article/us-usa-trump-congress-emergency-idUSKCN1QV1A8 |
| https://www.reuters.com/article/us-usa-politics-trump-idUSKCN1QH29L | https://www.reuters.com/article/us-usa-court-census-idUSKCN1QW2MC |

Query 5: Patriot Movement

|  |  |
| --- | --- |
| Lucene | PageRank |
| https://www.reuters.com/article/us-usa-trump-turkey-idUSKCN1QV328 | https://www.reuters.com/markets/bonds |
| https://www.reuters.com/article/us-france-protests-idUSKCN1Q325A | https://www.reuters.com/article/us-usa-election-orourke-idUSKCN1QV13C |
| https://www.reuters.com/article/uk-people-rkelly-metoo-idUKKCN1QC0V7 | https://www.reuters.com/article/canada-forex-idUSL1N2121OB |
| https://www.reuters.com/article/us-britain-climatechange-education-idUSKCN1QW2QU | https://www.reuters.com/article/us-climate-change-youth-idUSKCN1QW01S |
| https://www.reuters.com/article/us-climate-change-youth-idUSKCN1QW01S | https://www.reuters.com/article/us-slovakia-election-president-idUSKCN1QW2Z3 |
| https://www.reuters.com/article/us-climate-change-youth-idUSKCN1QW01S | https://www.reuters.com/article/us-climate-change-youth-usa-idUSKCN1QW1BE |
| https://www.reuters.com/article/us-climate-change-youth-idUSKCN1QW01S | https://www.reuters.com/article/us-climate-change-youth-usa-idUSKCN1QW1BE |
| https://www.reuters.com/article/us-italy-china-analysis-idUSKCN1QW1E2 | https://www.reuters.com/investigates/special-report/myanmar-suukyi-history/ |
| https://www.reuters.com/article/us-usa-election-orourke-idUSKCN1QV13C | https://www.reuters.com/article/us-auction-christies-georgemichael-art-idUSKCN1QP1EY |
| https://www.reuters.com/article/us-usa-election-orourke-idUSKCN1QV13C | https://www.reuters.com/article/us-italy-china-analysis-idUSKCN1QW1E2 |

Query 6: Oscar 2019

|  |  |
| --- | --- |
| Lucene | PageRank |
| https://www.reuters.com/article/us-awards-oscars-women-idUSKCN1Q7130 | https://www.reuters.com/ |
| https://www.reuters.com/article/us-awards-oscars-diversity-idUSKCN1QE0N4 | https://www.reuters.com/finance/markets |
| https://www.reuters.com/article/us-aviancaholdings-fleet-idUSKCN1QW2XT | https://www.reuters.com/politics |
| https://www.reuters.com/article/us-aviancaholdings-fleet-idUSKCN1QW2XT | https://www.reuters.com/news/world |
| https://www.reuters.com/article/us-aviancaholdings-fleet-idUSKCN1QW2XT | https://www.reuters.com/finance |
| https://www.reuters.com/journalists/pei-li | https://www.reuters.com/article/us-reuters-editorial-leadership-idUSKBN1KE2SD |
| https://www.reuters.com/article/us-britain-lgbt-celebrities-idUSKCN1QW2KY | https://www.reuters.com/video |
| https://www.reuters.com/theWire | https://www.reuters.com/video/breakingviews |
| https://www.reuters.com/article/us-britain-lgbt-rights-idUSKCN1QW2QI | https://www.reuters.com/investigates/section/homepage/ |
| https://www.reuters.com/article/us-britain-lgbt-rights-idUSKCN1QW2QI | https://www.reuters.com/finance/commodities |

Query 7: Channel

|  |  |
| --- | --- |
| Lucene | PageRank |
| <https://www.reuters.com/video/2019/03/11/in-brexits-next-act-theatrics-are-only-c?videoId=524541171&videoChannel=117760> | https://www.reuters.com/politics |
| https://www.reuters.com/video/2019/03/08/uk-japan-scientists-study-radioactive-fu?videoId=523497511&videoChannel=-10464 | https://www.reuters.com/article/us-reuters-editorial-leadership-idUSKBN1KE2SD |
| https://www.reuters.com/video/2019/03/13/woods-neck-not-painful-ahead-of-players?videoId=525458826&videoChannel=79 | https://www.reuters.com/video |
| https://www.reuters.com/video/2019/01/16/saudi-arabia-set-for-regions-largest-win?videoId=504158943&videoChannel=118285 | https://www.reuters.com/video/breakingviews |
| https://www.reuters.com/video/2019/03/11/tanzanian-mother-finds-fame-with-soccer?videoId=524707453&videoChannel=73 | https://www.reuters.com/markets/bonds |
| https://www.reuters.com/video/2019/03/15/malawi-floods-wreak-havoc-cyclone-idai-e?videoId=525916362&videoChannel=73 | https://www.reuters.com/subjects/supreme-court |
| https://www.reuters.com/video/2019/03/13/guardiolas-city-demolish-schalke-to-crui?videoId=525247223&videoChannel=79 | https://www.reuters.com/news/entertainment/arts |
| https://www.reuters.com/video/2019/03/04/breakingviews-tv-lyft-the-veil?videoId=521916948&videoChannel=117766 | https://www.reuters.com/news/us |
| https://www.reuters.com/video/2019/02/08/amazon-rethinks-ny-hq-over-opposition-re?videoId=512785961&videoChannel=5 | https://www.reuters.com/places/china |
| https://www.reuters.com/video/2019/02/01/jailed-reuters-reporters-appeal-to-myanm?videoId=510223083&videoChannel=-13280 | https://www.reuters.com/news/science |

Query 8: Wall

|  |  |
| --- | --- |
| Lucene | PageRank |
| <http://www.reuters.com/investigates/section/wall-streets-way/> | https://www.reuters.com/ |
| https://www.reuters.com/article/us-usa-trump-budget-wall-exclusive-idUSKBN1QR0CW | https://www.reuters.com/finance/markets |
| https://www.reuters.com/article/us-usa-immigration-wall-insight-idUSKCN1QW1GR | https://www.reuters.com/politics |
| https://www.reuters.com/politics | https://www.reuters.com/finance |
| https://www.reuters.com/article/us-usa-trump-congress-emergency-idUSKCN1QW2AK | https://www.reuters.com/video |
| https://www.reuters.com/article/us-usa-trump-congress-emergency-idUSKCN1QW2AK | https://www.reuters.com/video/breakingviews |
| https://www.reuters.com/article/us-usa-trump-congress-emergency-idUSKCN1QW2AK | https://www.reuters.com/investigates/section/homepage/ |
| https://www.reuters.com/article/us-usa-trump-congress-emergency-idUSKCN1QW2AK | https://www.reuters.com/finance/wealth |
| https://www.reuters.com/article/us-usa-trump-congress-emergency-idUSKCN1QV1A8 | https://www.reuters.com/finance/EarningsUS |
| https://www.reuters.com/journalists/noel-randewich | https://www.reuters.com/news/us |

|  |  |
| --- | --- |
| **Query** | **Overlap** |
| Venezuela | 0 |
| Senate | 2 |
| Democrats | 2 |
| Republicans | 3 |
| Patriot Movement | 3 |
| Oscar 2019 | 0 |
| Channel | 0 |
| Wall | 1 |